Retrieving Information Across Multiple, Related Domains Based on User Query and Feedback: Application to Patent Laws and Regulations

Hang Yu
College of Engineering
University of Illinois at Urbana-Champaign
Urbana, IL 60801, USA
hangyu@uiuc.edu

Siddharth Taduri
Engineering Informatics Group
Stanford University
Stanford, CA 94305-4020, USA
staduri@stanford.edu

Jay Kesan
College of Law
University of Illinois at Urbana-Champaign
Urbana, IL 60801, USA
kesan@law.illinois.edu

Gloria Lau
Engineering Informatics Group
Stanford University
Stanford, CA 94305-4020, USA
glau@stanford.edu

Kincho H. Law
Engineering Informatics Group
Stanford University
Stanford, CA 94305-4020, USA
law@stanford.edu

ABSTRACT

In this paper, we present a framework that can process user query for retrieval of information from documents of different properties across multiple domains, with specific application to patent laws and regulations.. A case example is given to demonstrate how results from multiple domain searches can be combined using ontology and cross referencing. A user feedback mechanism is also discussed in this paper.

Categories and Subject Descriptors

H.3.3 [Information Storage and Retrieval]: Information Search and Retrieval – retrieval models,

J.1 [Administrative Data Processing]: law.

General Terms

Algorithms, Design, Economics, Experimentation

Keywords

.patent, publication, search, ontology

1. INTRODUCTION

Huge amount of information is available on line and keeps increasing. More and more knowledge database is available on the Internet.

ICEGOV2010, October 25-28, 2010, Beijing, China

The fast pace of technological innovation is contributing to major changes in governments, societies, and the world economy. We are facing a problem of not being able to easily identify related documents across different information domains. A framework that can enable users to query multiple databases together would be desirable

Let us consider a few examples. If a company wanted to study the market for acid reflux drugs, they may choose to go to the FDA web site, they may look for court cases involving these drugs and they may also study some relevant technical publications. Similarly, a start-up company looking to work on therapeutics in the breast cancer space, may choose to study patents in this field, whether some patents were litigated, and the applicable scientific and technological literature.

In each situation, we have a common problem. There is relevant information that must be accessed and which is available in different information domains and the information is heavily soloed. In addition, even within one domain, the information may not be easily accessible and searchable. Broadly speaking, we have information on a particular topic in:

(a) an administrative agency;

(b) the court system;

(c) the relevant laws and regulations;

(d) other literature such as scientific publications.

Related to government regulations, here are administrative agencies that deal with various science and technology issues such as the Food and Drug Administration (FDA), the Environmental Protection Agency (EPA), the U.S. Patent and Trademark Office (USPTO), the Nuclear Regulatory Commission (NRC), the Federal

Communications Commission (FCC), and the like. The agencies promulgate regulations that appear in the relevant chapters of the Code of Federal Regulations (CFR) and they interpret these regulations and the applicable United States Code. In addition, the courts (often federal courts) interpret the relevant U.S. statutes and federal regulations (CFR). Moreover, there is often a need to consult additional literature in the form of technical/scientific publications. In general, for a given situation, such as evaluating the market and patentability of a new drug or technology invention, relevant information of different properties must be accessed; in practice, most relevant information do exist, and often accessible online nowadays but is available in different information domains and different formats and the information is heavily soloed.

In this paper we are focusing on a particular area of searching biotechnology patents. However, our research could be general enough to apply and adapt to other inter-agency information searches. We aim to build an information system for biotechnology patent management and related court litigations.

This paper discusses three basic components in our research and development efforts. The first is the creation of a document repository of core patents and publications using ontology. This repository includes a suite of concept hierarchies that enable users to browse documents according to the terms they contain. The second is an XML framework for representing documents features and associated metadata. The XML framework enables the augmentation of regulation text with tools and information that will help users understand and compare across prior published patents and publications. The third component is the creation of a feedback system with a user interface.

This paper is organized as follows: Section 2 will provide a background on our motivation and briefly review existing work in this area. Section 3 will describe the current online database we are using and our proposed framework. Section 4 will discuss a detailed user case of a well known biotechnology patents on erythropoietin (EPO). We will demonstrate how we can integrate patent searches and the scientific literature together. Section 5 will summarize and conclude this paper.

2. BACKGROUND

2.1 Motivation

With the advance of new biotechnology in the last decade, the number of biotech patent applications filed has soared. However, the tedious preparation of patent applications has become a burden for inventors and it has seriously undermined start-up and small business companies and inventors' efforts to protect their inventions.

The majority of the work in preparing a patent application has been spent in research related work and prior art. During the application process, inventors and patent lawyers want to answer the following questions:

(a) Are there any similar or related inventions that have already been patented ?

(b) Are there any similar or related inventions that are in the process of being patented ?

(c) Are there any similar or related inventions or techniques that have already been known or published ?

(d) Are they any similar or related inventions that are under being litigated in federal court?

To obtain the answers to all these questions, experienced engineers, patent agents and patent lawyers need to spend a lot of time researching various patent databases, academic journals and court documents, and tremendous efforts are made to cross reference each single document and centralize them.

Therefore, it is highly desirable for people to have one central place that they can obtain all sorts of documents in a well-classified form with good cross referencing notes. Although today's advanced information retrieval technique could lead to effective search for documents in a single domain, the capability to search multiple domains at the same time and centralize them in a well-sorted manner is still not achieved.

2.2 Our Goal

Our ultimate goal is to create a system that could help us to obtain related documents for a single search query across multiple domains. In this paper, we present some of our preliminary results on jointly searching USPTO patents and PUBMED scientific publications at the same time with good crossing referencing capabilities for bio-tech search terms.

2.3 Review of Previous Work

People have studied techniques in retrieval of scientific publications and patent searches. For instance, natural language processing techniques have been applied to search biomedical scientific publications [1] A two-stage retrieval method particularly using the claim structure has been proposed for patent searches [2].. However, most of the existing works focus on how to optimize and construct queries for document retrieval [3,4]. Some of them are also looking to automate their efforts. An example is to automatically generate patent search queries, as in [5]. Except for all these works, however, there is little effort to combine retrieval of documents from multiple, related

domains. Pioneer work in this direction originates from searching across multiple language versions of documents as in [6,7]. A few researchers have extended these efforts to multiple, loosely related domains like patents and news as in [8]. Although efforts have been made in patent retrieval by analysis of its citations as in [9], the effort aims at enhancing quality of patents retrieval instead of jointly searching both the patents and the scientific publications, as well as related documents and information..

3. SOLUTION FRAMEWORK

3.1 USPTO Patent Database

The United States Patent and Trademark Office (USPTO) web site provides free access to electronic copies of all existing patents, together with all materials in patent publications. The USPTO web user interface offers both quick and highly customized search for users. Certain analysis tools are also available. One disadvantage for USPTP website is that some of the documents are not in a searchable format. For instance, some documents exist in an image format (TIFF). Although third party vendors also exist to provide patents documents in other digital forms, as a first step, we focus on USPTO website as most issue patents can be accessed in text format. This is also the major approach that inventors use to search the USPTO's patent database to see if a similar patent has already been filed or granted. Patents may be searched in the USPTO Patent Full-Text and Image Database (PatFT). The USPTO houses full text for patents issued from 1976 to the present and TIFF images for all patents from 1790 to the present.

3.2 NIH Scientific Publication Database

The Entrez Global Query Cross-Database Search System is a powerful search engine with a web user interface, through which users can search multiple databases hosted at the National Center for Biotechnology Information (NCBI) website in health science. NCBI is part of the National Library of Medicine (NLM), itself a department of the National Institutes of Health (NIH) of the United States. Entrez global query system is a search and retrieval system which links to multiple databases. It can access all these databases at the same time with one user query. It also has a unified user interface. Besides scientific publications, it also contains related data like DNA sequences and structures. As there are several databases in Entrez, we pick Medical Literature Analysis and Retrieval System Online (MEDLINE), which can be accessed via PUBMED, as our gateway to Entrez MEDLINE is a bibliographic database of life sciences and biomedical information. We can not only find most bibliographic information for articles from areas like medicine, nursing, pharmacy, dentistry, veterinary medicine, and health care but can also find much of the literature in general biology and biochemistry, as well.

Areas like molecular evolution are also included. In other words, this is a very complete database, in which we expect to find most of the scientific publications that people refer to in the biotechnology and biomedical areas.

3.3 Framework of Joint Search

Our joint search system has three components. The first component is ontology mapping and generation. What happens is that the keywords entered by users are mapped into a subset of relevant keywords. This step is performed by looking those words up in ontology database. The second component is the joint and cross search in various documents domains; in our case, they are patents and scientific publications. As our goal is to support joint search in multiple domains, those databases can well be located in the Internet/WWW instead of being saved locally. As an example, we could use a computer script to automatically search USPTO website to look for patents that are most relevant with these keywords. These patents would be considered as "core" patents. Next, we extract all scientific publications cited by these core patents and apply cross referencing analysis on them. The last component is to modify the search results by applying user feedback statistics. The results of feedback will be saved as meta data for future uses.

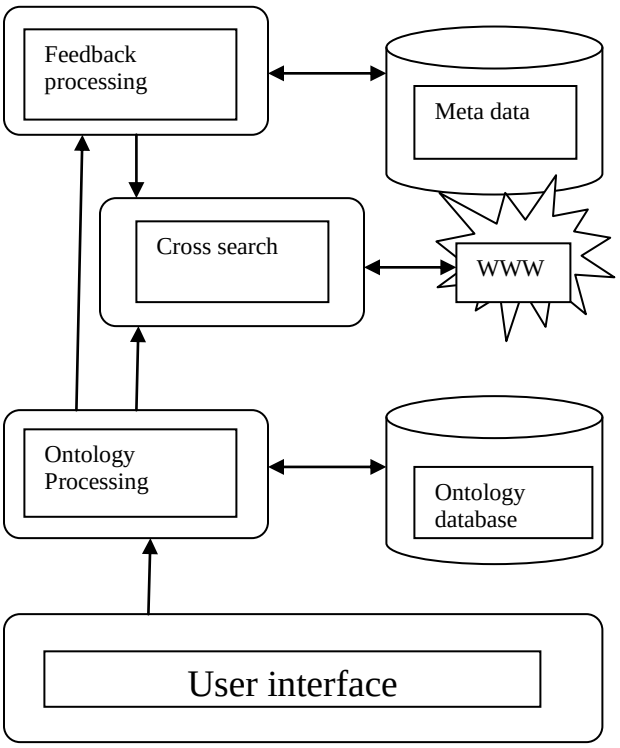

Figure 1: System Framework

3.4 User feedback

The ability to take user feedback into the framework is important. There is no doubt that domain knowledge from expert or experience users could be a very good compliment to our system.

User feedback could exist in two forms: indirect and direct. Direct forms are feedbacks that are immediately obtained at the user interface level. Users can enter feedbacks by click buttons or enter values in a user interface, either on a web page or an application. Indirect forms of feedback are those implicitly expressed by the users. Typical examples include number of citations by other documents or number of queries a system received. Google's page rank is a good example of indirect user feedback as web pages receive their feedbacks by the number of other pages that have links to them. In our case, we could use the number of citations a publication gets from a subset of related patents as its indirect user feedback.

User feedback could also be modeled by different approaches depending on the real interface we have. For instance, users can be asked to rate a publication that appear for his/her query in numerical form (scores from 1 to 5) or users can be asked to give binary (click "positive" if satisfied) feedbacks or tertiary ones ("positive", "neutral", "negative"). In this research, we will demonstrate the simple usage of binary feedbacks:

(a) In direct form, the user would be able to click a button to express that he/she is satisfied with the result

(b) In indirect form, the user would express his/her satisfaction by citing a publication in his/her patent application. We propose an algorithm that would assume that user feedback is always correct to the best of the user's knowledge (in other words, we assume user enter feedback in good faith).

We will always include a publication if user feedback score is larger than a threshold (**TH**) before we use our normal procedure to determine if a publication is relevant or not. The raw feedback score (**Rufs**) is an aggregate "positive" feedback normalized by the total number of visits a document has.

$$Rufs(i) = \frac{number\ of\ user\ feedbacks}{number\ of\ user\ visits}, for\ doc\ i$$

Simply applying this formula could obviously be biased by the user's habit of leaving a feedback since not all users leave feedback. Some users are more active and some are not. To minimize the bias, the Rufs is adjusted by the average user feedback ratio (**Aufs**):

$$Aufs = \frac{total\ number\ of\ user\ feedbacks}{total\ number\ of\ user\ visits}, for\ all\ docs$$

Thus a final feedback score **(Fufs**) could be defined as:

$$Fusf = \frac{Rufs(i)}{Aufs}$$

The acceptance rule could be defined as:

Accept if Fufs(i) >= TH

where TH is a threshold value the system uses to reflect its belief on the experience level of the users. For general users, we can set the threshold to be high to consider those documents highly recommended by the users. For a system that is used by extreme experience users, we can lower this threshold to rely on more expert feedbacks. In an extreme case when TH goes to zero, the system would include search results if any one of the expert users has recommended it. This model could be easily extended to take into consideration of users with differently levels of experience by weighting their opinions. However, as a first step, we would assume all users are of the same level of experience when providing their feedbacks

4. AN EXAMPLE CASE STUDY: JOINT SEARCH OF ERYTHROPOIETIN (EPO)

4.1 Background

Erythropoietin is a glycoprotein hormone that controls erythropoiesis, or red blood cell production. It is a cytokine for erythrocyte (red blood cell) precursors in the bone marrow. EPO is also produced by the peritubular capillary endothelial cells in the kidney, and is the hormone that regulates red blood cell p reduction. In 1968, Goldwasser and Kung began work to purify human EPO, and managed to purify 10 ml by 1977, nine years later. The pure EPO allows the amino acid sequence to be partially identified and the gene to be isolated. Later, an NIH-funded researcher at Columbia University discovered a way to synthesize it. Columbia University patented the technique and licensed it to Amgen.

Amgen later had patents based on innovations made by its scientist, Dr. Fu-Kuen Lin, related to a naturally occurring human hormone called erythropoietin, or EPO, that stimulates the production of oxygen-carrying red blood cells. When Swiss drug maker Roche sold its anemia drug Mircera in the United States to compete with Amgen's rival drugs companies Aranesp and Epogen, Amgen filed a lawsuit. Roche's main counter argument is that Amgen's patents are not valid because the technology underlying production of the drugs was already in the public domain before Amgen filed for patent protection in 1984.

In our research, we would use this case example to demonstrate how a joint search framework could lead us to patents and original academic publications.

4.2 Ontology

As bio-tech search terms are mostly strict scientific terms and may have different meaning in different domains, we have to first establish a mapping across multiple domains to make sure we mean the same thing for each of them. Generally, this is achieved by establishing an ontology and we could generate a list of related terms from a single term.

We obtain a list of related words by using the ontology founded at Bio Portal. As an example, Figure 2 is a list of key words obtained by using various ontology databases for "EPO".

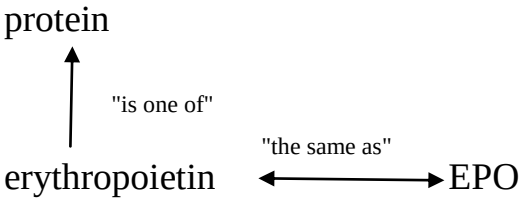

Figure 2 Example for an ontology

4.3 Core Patents

Table 1 shows the five core patents of EPO litigations. To come up with these 5 patents, we first search in USPTO with a query using keywords and obtain a large set of relevant patents. We then pick these five most important patents, identified by reading several court cases and consulting with several experienced patent litigators, as core patents. They appear the most number of times in the original lawsuits.

Table 1 List of core patents for EPO

U.S. Patent Number	Date
5,621,080	04/15/1997
5,756,349	05/26/1998
5,955,422	09/21/1999
5,547,933	08/20/1996
5,618,698	04/08/1997

4.4 Core Publications

After the five core patents are identified, we manually extracted all publications cited by these patents to establish a database based on those publications. The 300 publications extracted are considered as the core publications. To give an example, we list a few examples. in Table 2 (where PUBMED Id is the index PUBMED gives to each publication.)

Table 2 List of some core publications related to core patents

PUBMED ID.	Title	Referenced In
6713094	Evidence for the Presence of CFU-E with Increased In Vitro Sensitivity to Erythropoietin in Sickle Cell Anemia	5621080, 5756349, 5955422, 5547933, 5618698
3680293	Structural Characterization of Natural Human Urinary and Recombinant DNA-derived Erythropoietin	5621080, 5955422, 5547933, 5618698
3624248	Carbohydrate Structure of Erythropoietin	5621080,

	Expressed in Chinese Hamster Ovary Cells by a Human Erythropoietin cDNA	5756349, 5618698
232226	Cloning of Hormone Genes from a Mixture of cDNA Molecules	5955422, 5547933
14025852	Current Concepts in Erythropoiesis	5547933

4.5 Extracting Features on Publications

To determine if a scientific publication is important or relevant to a patent, we need to extract certain features and quantify them. In this work, we would use the word frequency of a particular key word in a scientific publication's abstract. Here is an example, as show in Table 3

The original abstract is shown in Figure 3 where the key terms are underlined. Note the key term appears 5 times and the total word count is 159 therefore the keyword term frequency, as tabulated in Table 3, is:

$$\frac{5}{159} = 3.145\%$$

Table 3 Some feature for a selected publication

Title	Human erythropoietin gene: High level expression in stably transected mammalian cells and chromosome localization
Author	Powell et al.
Journal	P.N.A.S. (USA), 83, 6465-6469 (Sep. 1986).
Keyword	erythropoietin
Keyword Count	5
Abstract Word Count	159
Keyword Frequency	3.145%

Original Abstract *The glycoprotein hormone erythropoietin plays a major role in regulating erythropoiesis and deficiencies of erythropoietin result in anemia. Detailed studies of the hormone and attempts at replacement therapy have been difficult due to the scarcity of purified material. We used a cloned human erythropoietin gene to develop stably transfected mammalian cell lines that secrete large amounts of the hormone with potent biological activity. These cell lines were produced by cotransfection of mammalian cells with a plasmid containing a selectable marker and plasmid constructions containing a cloned human erythropoietin gene inserted next to a strong promoter. The protein secreted by these cells stimulated the proliferation and differentiation of erythroid progenitor cells and, with increased selection, several of these cell lines secrete up to 80 mg of the protein per liter of supernatant. Hybridization analysis of DNA from human chromosomes isolated by high resolution dual laser sorting provides evidence that the gene for human erythropoietin is located on human chromosome 7*

Figure 3 Original abstract

Besides the key term frequency in the abstract, we have also included multiple other features that are useful. Examples include the key term's appearance in the title, the key terms' appearance in the full text of the article. In this paper, we focus on key term frequency in the abstract.

Applying this analysis to all the patents, we can obtain a XML file as shown in Figure 4. It should be noted that the full text and some features are not presented in the figure due to space limitation.

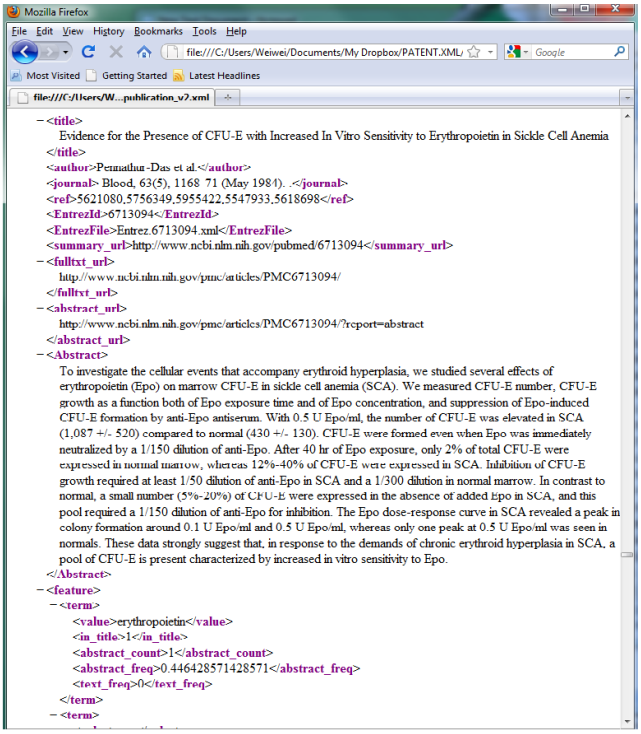

Figure 4: XML file to contain feature and metadata

To summarize, we present some results on some key search words in terms of frequency in the abstract as shown in Table 4. Note RefScore will be explained below and "erythropoietin", "EPO" and "protein" are three keywords obtained by an ontology lookup.

Table 4 Word frequency for some sample key words

Paper Id	Ref Score	**Erythro-poietin**	**EPO**	**protein**
6713094	5	0.446	7.59	0
2813359	5	1.093	8.74	0
18202227	5	0.565	3.96	0.565
3680293	4	0.467	3.74	1.402
3624248	3	3.265	0	1.224
232226	2	0	0	0
14025852	1	0	0.	0

4.6 Examining Correlation between Features and Relevance

First, we need to define the "relevance" of a publication for a search term. We use a simple metric **RefScore** to quantify each document's relevance to the search term: RefScore is defined as the number of times a publication is cited by the five core patents. In this case study, it has max value of 5 since there are five core patents. A scientific publication that has a RefScore of 5 is cited by all of our core patents so it is reasonable to consider it as one of the most relevant publication.

Once we have established the relevance, we further analyze the statistic relationship between RefScore and key term frequency. In the graph shown in Figure 5, the X-axis is the number of word frequency in the abstract and the Y-axis is the Ref Score. We can see that high number of frequency is a good indicator of good Ref Score. When it exceeds a certain level, all sample publications have a high Ref Score of 5.

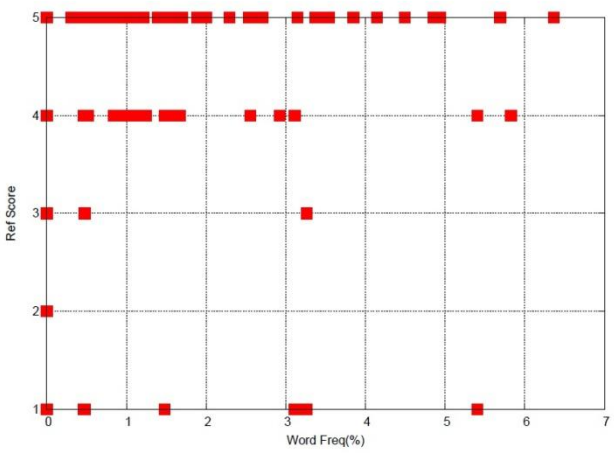

Figure 5 Relationship between RefScore and key term frequency

To further extend this analysis to all documents, we calculate the correlation coefficient of RefScore and key term frequency. The results are shown in Table 5. A positive value indicates that the feature is positively related to RefScore and therefore is a valid feature. In this table, we could see that those key words listed in the right column are good indicators whether a publication would be used in related patents applications.

Table 5 Correlation between keywords and RefScore

Keyword	Correlation
Erythropoietin	0.089
Epo	0.08
Iron	0.065
Erythropoietin	0.035
Cytokines	0.035
Desamethasone	0.035
hydroxyurea	0.035
Protein	-0.002

4.7 User Feedback via Citation and/or Direct Interactive Interface

As direct interactive form user feedback is not easily obtained, we present an example of indirect user feedback. We use the number of times a publication is cited in a core patent as a "positive" vote and use the number of citations in PUBMED to measure the visits a publication could get. In this case, a user feedback score (Rufs) in this case is defined as

$$Rufs(i) = \frac{number\ of\ citations\ in\ patents}{number\ of\ citations\ in\ PUBMED}$$

And we can define the normalization factor or the average user feedback score (**Aufs**) as:

$$Aufs = \frac{total\ number\ of\ citations\ in\ patents}{total\ number\ of\ citations\ in\ PUBMED}$$

Finally, we can define the final user feedback score (**Fufs**) as:

$$Fufs(i) = \frac{Rufs(i)}{Aufs}$$

Using the above seven publications as an example, the scores for all the seven publications are obtained as shown in Table 6:

Table 6 A user feedback example

Paper	RefScore	#Citation	Rufs(%)	Fufs
6713094	5	219	2.28	0.94
2813359	5	134	3.73	1.54
18202227	5	260	1.92	0.79
3680293	4	119	3.36	1.38
362424	3	98	3.06	1.26
232226	2	103	1.94	0.80
14205852	1	98	1.02	0.42
Total	25	1031	2.42	--

If we set a threshold of TH = 1.50, we could see that publication 2813359 meets the threshold. Therefore, we consider this document to be highly applauded by the users and it will always come up in the subsequent searches.

We should also note that the user feedback is a dynamic process. With the passing of time, if publication 2813359 is cited by more PUBMED publications and becomes less cited in new patents, it will eventually be removed from the category of "positive feedback".

4.8 User Interface

Here, we present an example of the user interface page in Figure 6. The search results are listed item by item. Under each item, we have established two links to the PUBMED web page for the user to access the original paper summary and download the full text PDF file. We also add a "Vote Positive" link for each publication for users to enter feedback. If a user clicks this link, meta data for this publication will be adjusted and saved in the backend database.

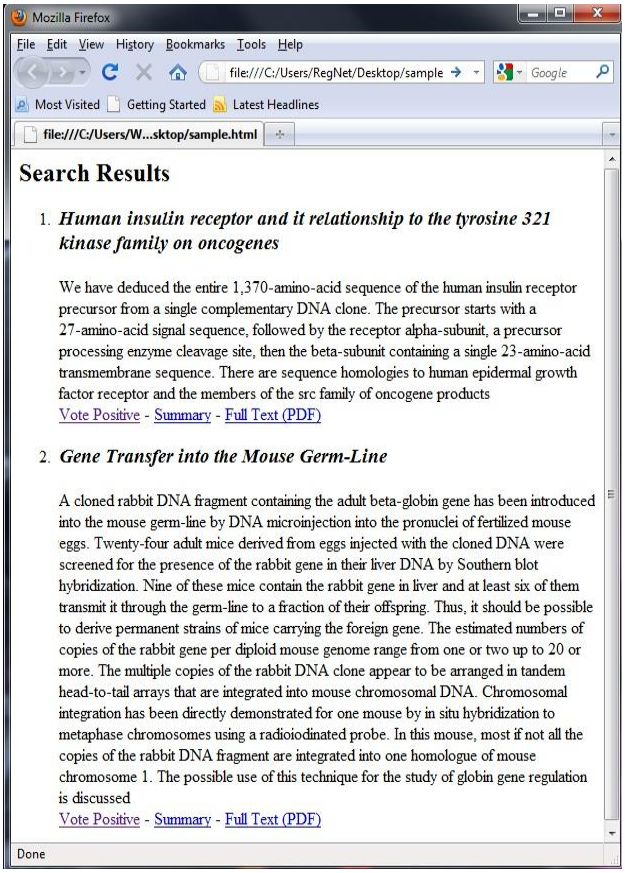

Figure 6 User interface example webpage

5. SUMMARY

In this paper, we have presented a framework to jointly search patents and scientific documentations. However, our endeavor to achieve a universal integrated search system doesn't end here. Further extensions could be made to our system. Besides issued patents and scientific documents, there are many related document resources that are available. For instance, USPTO office offers all patent application prosecution materials for all issued patents, commonly referred to as the File Wrapper. These documents provide more details on the searches and analysis conducted by the USPTO examiners and the patentees' responses for all patent applications and the relevant prior art. Another example is that all patent litigation cases could also be found on line on the PACER website for all U.S. federal district and appellate courts. These court documents could also be interesting to the users who are examining particular document or researching a particular technology product.

Our method and framework presented in this paper could be easily extended to include more than two document domains. We could include more heterogeneous domain knowledge so that the search results could be more relevant

and robust. It is more likely that if a document is cross-referenced in four domains is more important and relevant than the documents only appears in two domains. In future study, we will focus on a more generic framework that could include more types of document databases.

6. ACKNOWLEDGEMENTS

This research is partially supported by NSF Grant Number 0811975 awarded to the University of Illinois at Urbana-Champaign and NSF Grant Number 0811460 to Stanford University. Any opinions and findings are those of the authors, and do not necessarily reflect the views of the National Science Foundation.

7. REFERENCES

[1] L Hunter, KB Cohen, Biomedical language processing: what's beyond PubMed, *Molecular Cell, 2006 - Elsevier*

[2] Hisao Mase et al , Proposal of two-stage patent retrieval method considering the claim structure, , *ACM Transactions on Asian Language Information Processing (TALIP) archive, Volume 4 , Issue 2, June 2005, Pages: 190 - 206*

[3] H Tseng, CJ Lin, YI Lin, Text mining techniques for patent analysis, , *Information Processing & Management, 2007, Elsevier*

[4] T Takaki , Associative document retrieval by query subtopic analysis and its application to invalidity patent search, Conference on Information and Knowledge Management, *Proceedings of the thirteenth ACM international conference on Information and knowledge management table of contents, Washington, D.C., USA, Pages: 399 - 405, 2004*

[5] Xiaobing Xue et a, Automatic query generation for patent search Conference on Information and Knowledge Management, *Proceeding of the 18th ACM conference on Information and knowledge management, Hong Kong, China*

[6] Y Li, J Shawe-Taylor , Advanced learning algorithms for cross-language patent retrieval and classification , *Information Processing & Management, 2007, Elsevier*

[7] PRIME: A system for multi-lingual patent retrieval, , *Proceedings of MT summit VIII, 2001*

[8] S Higuchi, M Fukui, A Fujii, T Ishikawa, M Iwayama, A Fujii, N Kando, An empirical study on retrieval models for different document genres: patents and newspaper articles, *Information retrieval, 2003*

[9] A Fujii, Enhancing patent retrieval by citation analysis, *Annual ACM Conference on Research and Development in Information Retrieval, Proceedings of the 30th annual International ACM SIGIR, Amsterdam, The Netherlands*

[10] http://bioportal.org/, June 04, 2010